Data Science Project Proposal
Firas Mouasher, MSIM'19

Executive Summary:

As suggested by my chosen dataset's title, "New York City Major Felony Incidents" provides various data on major felonies that took place in New York City, at the incident level. I plan on focusing on providing the New York Police Department (NYPD) with information that could help them better allocate how many officers should be in each respective jurisdiction at a given time. Furthermore, I hope that I could provide the NYPD with information on the different types of crimes that could occur at any given time and place. This will be beneficial as different felonies require different measures and safety precautions to be taken. For example, some of the major felonies listed include, but are not limited to murder and non-negligent manslaughter, rape, robbery, felony assault, and burglary.

The 'borough' field will be an independent variable so that we only focus on Manhattan and its respective precincts. First, some of the important dependent variables include the year, month, day, and hour fields. Ideally, I would like to measure the frequency of felonies committed in Manhattan with emphasis on different time frames in order to better understand in which days, months or seasons felonies of which type are to occur. Another couple of important dependent variables are both the 'offence' and the 'precinct' fields. The type of offence and frequency of the offences across the aforementioned time frames are important to keep in mind so that the NYPD are able to understand for which crimes to better prepare for; a high frequency of shootings on a given day may prompt armed police forces while a high probability of theft would require more oversight and added precautionary measures. The 'precinct' field will also play an integral role in that it will be able to divide Manhattan into thirty-four precincts, which could be used to predict where the crimes are most likely to happen within Manhattan.

Currently, the dataset I will be using will need to be cleaned and some issues need to be addressed. The dataset includes 1,123,465 rows without holding the 'borough' variable constant. When the borough is made an independent variable in order to focus solely on Manhattan, the number of rows reduces to 287,812 rows. The datatypes included in this dataset include strings, datetime, and integers. The dataset also includes some unnecessary fields and variables that will not necessarily be used such as the 'X-Coordinate' and 'Y-Coordinate' fields that attempt to over-specify the location in which the felonies occur per respective precinct. In addition, I will need to make sure that all fields carry appropriate and consistent data types and that null values and errors do not exist after processing. Furthermore, some of the felonies committed could be not valuable and make the data more noisy, messy, and difficult to interpret. In those cases, I will try to focus on noticeable trends that prompt obvious actionable measures to be taken. Other more irrelevant felonies could be cleaned from the dataset.

Given the dependent variables mentioned above, I aim to present a tool to the NYPD that will allow them to better allocate their resources across each precinct according to the probability of felonies occurring across time and precincts. Ideally, I would like to present the data in a clear and useful way using an interactive map of Manhattan that is divided up into precincts as shown in the map below. I aim to gain insights into some of the many challenges the

Data Science Project Proposal
Firas Mouasher, MSIM'19

Federal Investigation Borough (FBI) faces when predicting how to better allocate the number of officers that should be in each respective jurisdiction at a given time- frame.

The data is collected and provided by the NYPD, and due to obvious security reasons, the NYPD deliberately withholds from publishing detailed and updated methodologies explaining their data collection methods, process, storing, and publishing. This particular dataset provides data for felonies that had occurred between 2006 and 2010. Ideally, a more updated dataset in order to extract insights that could predict the future in order to be used to aid the NYPD better allocate their resources with respect to time and location. In addition, the number of data points decreases every year from 2006 through 2010 (128,441 data points in 2006 and 105,636 data points in 2010). It is unclear whether the decrease in the number of data points provided is attributed to a steady decrease in crime rate during the aforementioned period. Furthermore, for privacy reasons, incidents have been moved to the midpoint of the street section on which the respective felonies occur.

The dataset described above is owned by NYC Open Data by the city of New York, found here: https://data.cityofnewyork.us/Public-Safety/NYPD-7-Major-Felony-Incidents/hyij-8hr7/data. In March 2012, former May Bloomberg signed Local Law 11 of 2012, which is described as "Open Data Law", which required all public data be made available on a single web portal by the end of 2018 to provide greater visibility to community stakeholders. It is important to understand how NYC Open Data became what it is today and its impact on ethics and overall public interest.

The notion that government activities and data should be "open" date back to NYC in the early 20th century, when Progressive Era reformers "fought for legal remedies to the corrupt ward politics and favors-trading power brokerage characterized by the Tammany Hall political machine" (https://www.govtech.com/data/New-York-City-Open-Data-A-Brief-History.html). These reformers sought to ensure public officials acted in the best interest of the public. In 1993, the NYC Commission on Public Information and Communication (COPIC) created a Pubic Data Directory in order to better manage information and make it more accessible to the public by leveraging improvements in technology and the World Wide Web. Since then, Open Data directives continue to be important, even at the Federal level, when Barback Obama issued the "Memorandum on Transparency and Open Government" on his first day as president in 2009 which sought to establish a more transparent and collaborative government through an open data policy titled, the Open Government Directive. The promise and subsequent execution of an open data platform that was the basis for the Enigma distribution is one step towards shedding light on government activities in an effort to provide transparency to community stakeholders, mitigate potential corruption, and inspires others to act on their own accord.

While open data provides the aforementioned benefits, opening data may also have negative or unexpected consequences by providing data access that was previously difficult or not possible to find. If NYC focuses too much on opening up the data but not on the intended use of that data, then they could potentially undermine the purpose of providing NYC Open Data in the first place. In "The ethics of big data as a public good: which public? Whose good?", author Linnet

Data Science Project Proposal
Firas Mouasher, MSIM'19

Taylor discusses how creating access to datasets would distribute some of the power held by firms or governments to researchers. Further, as more and more decision makers are involved in data analysis, the power would also shift away from the individuals who originally created the data and who are most likely to be affected by potential misuse.

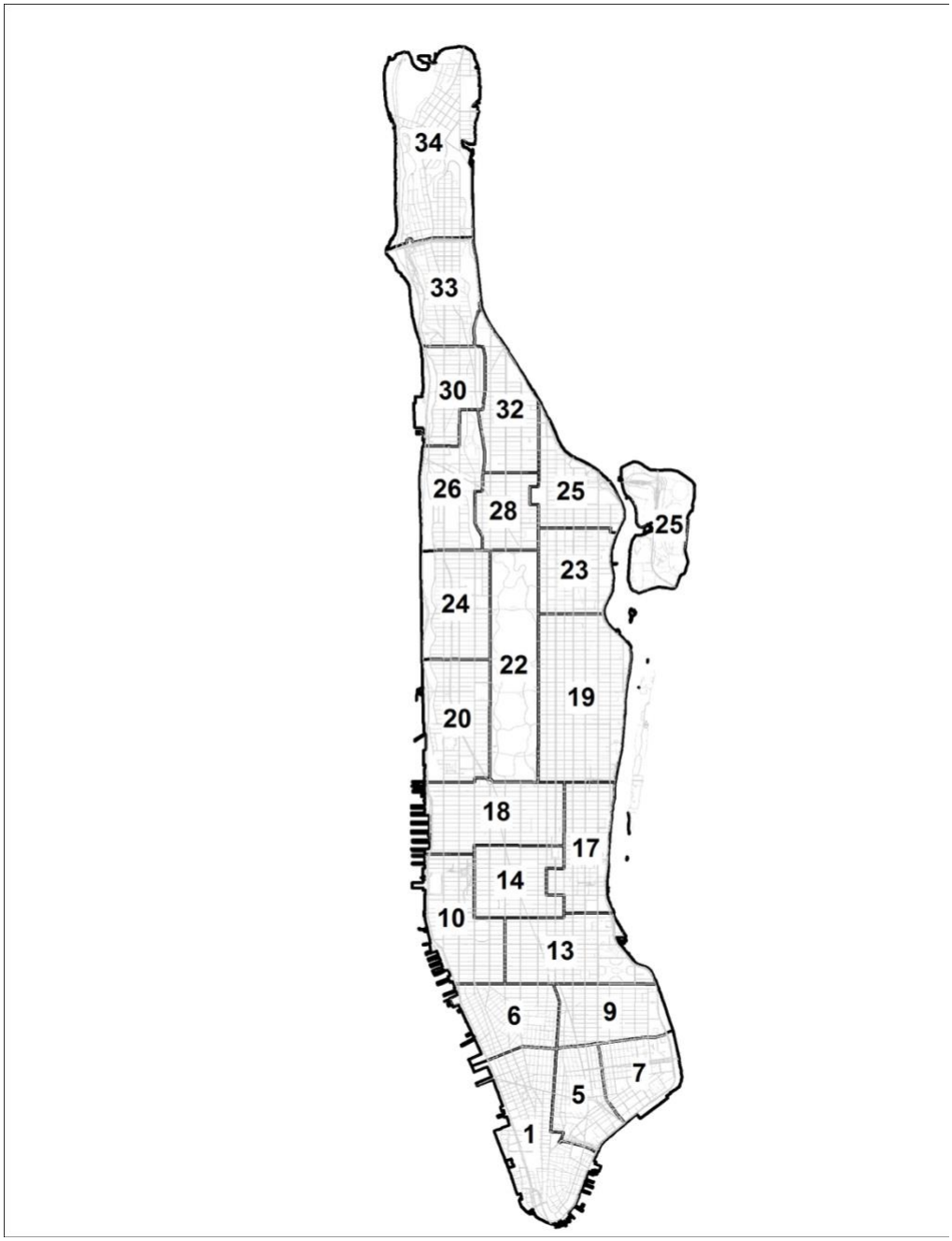
In the world of research ethics today, ethics and privacy need to be considerations for research, and those decisions should occur in public and remain part of the public record. "Open ethics" in regards to what research should be conducted in the first place should be just as important as NYC's push for "open access" and "open data".

Additional footnotes provided by the NYPD regarding data collection and interpretation:
Source: NYPD Arrest Incident Level Data Footnotes

- Information is accurate as of the date it was queried from the system of record, but should be considered a close approximation of current records, due to arrest revisions and updates.
- Data is available as of the date that technological enhancements to information systems allowed for data capture. Null values appearing frequently in certain fields may be attributed to changes on official department forms where data was previously not collected. Null values may also appear in instances where information was not available or unknown at the time of the report and should be considered as either "Unknown/Not Available/Not Reported."
- Arrests occurring near an intersection are represented by the X coordinate and Y coordinates of the intersection. Arrests occurring anywhere other than at an intersection are geo-located to the middle of the nearest street segment where appropriate.
- Any attempt to match the approximate location of the incident to an exact address or link to other datasets is not recommended.
- Many other arrests that were not able to be geo-coded (for example, due to an invalid address) have been located as occurring at the police station house within the precinct of occurrence.
- Arrests occurring in open areas such as parks or beaches may be geo-coded as occurring on streets or intersections bordering the area.
- Arrests occurring on a moving train on transit systems are geo-coded as occurring at the train's next stop (street intersection).
- X and Y Coordinates are in NAD 1983 State Plane New York Long Island Zone Feet (EPSG 2263).
- Latitude and Longitude Coordinates are provided in Global Coordinate System WGS 1984 decimal degrees (EPSG 4326).

Link to Dataset I will be using:
https://public.enigma.com/datasets/new-york-city-major-felony-incidents/9bff20f8-4476-4f20-9562-5c608872fadc

Data Science Project Proposal
Firas Mouasher, MSIM'19



Data Science Project Proposal
Firas Mouasher, MSIM'19

Sources:

1. file:///Users/firasmouasher/Downloads/NYPD_Arrest_Incident_Level_Data_Footnotes.pdf
2. https://public.enigma.com/datasets/new-york-city-major-felony-incidents/9bff20f8-4476-4f20-9562-5c608872fadc